

Method of determining an eigenspace for representing a plurality of training speakers

The invention relates to a method of determining an eigenspace for representing a plurality of training speakers. With such a method, initially speaker-dependent sets of models are developed for the individual training speakers while training speech data of the individual training speakers are used and the models of a set of models are described each time by a plurality of model parameters. In a high-dimensional vector space a combined model for each speaker is then displayed, in that a plurality of model parameters of the sets of models of the individual training speakers is concatenated to a respective coherent supervector. Subsequently, when the dimension is reduced, for realizing basis vectors for the eigenspace, a transformation to the supervectors of the training speakers is then performed. In addition, the invention relates to a method of speech recognition in which a basic set of models is adapted to a current speaker while the eigenspace generated according to the invention is used.

Speech recognition systems usually work in the way that first the speech signal is spectrally or chronologically analyzed in an attribute analysis unit. In this attribute analysis unit the speech signals are usually divided into sections, so-called frames. These frames are then coded and digitized in suitable form for the further analysis. An observed signal may then be described by a plurality of different attributes or, in a multidimensional attribute space, by an "observation" vector or "attribute" vector. The actual speech recognition, i.e. the recognition of the semantic content of the speech signal, finally takes place in that the sections of the speech signal described by the observation vectors or by a whole sequence of observation vectors, respectively, are compared with models of different practically possible sequences of observations and in this way a model is selected that matches the observation vector or sequence found best. The speech recognition system is therefore to comprise a kind of library of all possible signal sequences from which the speech recognition system can then select the respectively matching signal sequence. That is to say, the speech recognition system contains a set of acoustic models for different acoustic units which could, in principle, practically occur in a speech signal. The acoustic units may be, for example, phonemes or phoneme-like units such as diphones or triphones, in which the model of the phoneme depends on the previous and/or subsequent phoneme within a context.

Obviously, the acoustic units may also be complete words. Accordingly, such a set of models may consist of only models of phonemes, diphones, triphones or the like, of words or a mixture of different acoustic units.

A method often used for describing acoustic units i.e. certain sequences of observation vectors, is the use of so-called Hidden Markow models (HM-models). They are stochastic signal models for which it is assumed that a signal sequence is based on a so-called "Markow chain" of different states, where certain transition probabilities exist between the individual states. The respective states themselves cannot be recognized then (hidden) and the occurrence of the actual observations in the individual states is described by a probability density in dependence on the respective state. A model for a certain sequence of observations may therefore be described in this concept by the sequence of the various transited states, by the duration of the stop in the respective states, the transition probability between the states and the probability of occurrence of the individual observations in the respective states. A model for a certain phoneme is then generated, so that first suitable start parameters are used for a model and then, in a so-called training of this model by a change, the parameter is adapted to the phoneme to be modeled of the respective language until an optimal model has been found. The details about the various HM-models, as well as the individual exact parameters to be adapted, do not play an essential role in the present invention and are therefore only described hereinafter in so far as they are necessary for providing a more complete understanding of the invention.

Models may be either speaker-dependent (so-called SD models) or speaker-independent (SI models). Speaker-dependent models are trained exactly to an individual speaker in that the speech recognition system is supplied with a large number of examples of words of the individual speaker beforehand, while the contents of these word examples are known to the system. These examples are called training data. Such speaker-dependent systems or models, respectively, are relatively exact for the respective individual for which it has been trained. However, they are extremely inexact for any other person. With a speaker-independent system, on the other hand, the training of the models takes place based on training data of a large variety of different speakers, to thus generate models with which the system is in a position to understand any person that speaks the respective language. The error rate in such a speaker-independent system when speech data are to be recognized of a speaker who does not belong to the training speakers, however, is about 20 to 25% higher than the error rate for a comparable speaker-dependent system, which was trained specifically for this speaker.

In many applications of speech recognition systems, for example, when used in automatic telephone answering systems, there is unfortunately no possibility of training the system or models, respectively, to a certain speaker beforehand. To improve the performance of such systems, many speech recognition systems have meanwhile the possibility of adapting the system to the speaker during speech recognition, on the basis of the speech data already observed. A simple possibility of this is to transform the observations to be recognized coming from the current speaker, so that they are closer to the observations of a reference speaker for which a speaker-dependent model was trained. A further possibility is to group the training speakers according to their similarity and train a common model for each group of similar speakers. For the respective current speaker the model of the group is then selected which the speaker fits in best. A further, very good and effective method is the adaptation of a basic HM model to the respective speaker i.e. various HM parameters are changed, so that the changed model better matches the respective speaker. With this method a speaker-dependent model is formed during the speech recognition. The two best known model adaptation techniques are the so-called Maximum a Posteriori estimation (MAP) and the Maximum Likelihood Linear Regression method (MLLR). These highly effective techniques, however, have the disadvantage that they need both considerable computing power and time as well as a sufficient number of spoken signals of the new speaker until the suitably formed model is adapted to this new speaker.

EP 0 984 429 A2 therefore proposes a new type of method of adapting a speech recognition system to a speaker to be recognized, which system works with the so-called eigenvoice technique. A starting point of this eigenvoice method is the representation of speakers and their combined acoustic models as elements of a high-dimensional linear space, in the following to be referred to as model space, in which all the parameters describing a speaker are concatenated to a "supervector". A linear transformation is then performed with these supervectors of the training speakers, by which transformation the eigenspace basis vectors for the so-called eigenspace are recovered. This eigenspace is a linear subspace of the high-dimensional model space. The transformation is then performed in the way that the eigenspace basis vectors represent various correlation or discrimination attributes between the various training speakers or models, respectively, of the training speakers. A possibility of the transformation is mentioned herein, for example, the Principal Component Analysis (PCA) in which a correlation matrix is formed by means of the

supervectors of the various speakers and the eigenvectors of this correlation matrix are determined to be eigenspace basis vectors. Further possible methods are the "Linear Discriminant Analysis"(LDA), the "Factor Analysis"(FA), the "Independent Component Analysis"(ICA) or the Singular Value Decomposition (SVD). All these transformations are, however, relatively intensive as regards computations. Each of the so-provided eigenspace basis vectors represents a different dimension in which individual speakers can be distinguished from each other. Furthermore, based on the original training material each supervector of each speaker can be described by a linear combination of these basis vectors.

A problem of the eigenvoice method, however, is found in the conversion of this method to be used in the recognition of continuous speech with a large vocabulary. With such a speech recognition there are considerably more possibilities of successive phonemes and there is more often a wear of syllables etc. than when individual fixed commands are given. The real distribution of the observations is therefore too diffuse and there are too many variations. When HM-models are used, for example, an acoustic unit can no longer be described by a single state or by a plurality of separate states which are described by only a single probability density, for example, a single Gaussian or Laplace density. Instead, a mixture of various densities is necessary i.e. a plurality of such densities having different weights have to be superimposed to reach a probability density adapted to the real distribution.

If a language that has 42 different phonemes is started from, and if each of these phonemes is described by only three states per phoneme (initial state, middle state, end state), this will already lead to 142 different states which are to be described. When context-dependent phonemes are used, which is very practical when continuous speech is recognized, various context-dependent models are trained for each phoneme, depending on which phoneme immediately precedes and/or immediately succeeds (triphone). For describing such triphones of a language, a total of, for example, 2000 states are necessary. When a sufficient number of different probability densities per state (about 30) are used, there are about 60,000 different probability densities. With the customarily used attribute space of about 30 to 40 dimensions, this leads to a single speaker being described in the end by approximately two million individual model parameters. These model parameters comprise all the attribute parameters for describing 60,000 probability densities in the attribute space while, as a rule, only the mean value of each density is laid down in the attribute space and the variance for all the densities is assumed to be the same and constant. Obviously, for each density also additional parameters can be used, which individually determine the covariance for this

density. Besides, the model parameters may comprise, for example, the transition probabilities between the states and further parameters for describing the various HM models. The approximately two million model parameters are then to be concatenated to the supervectors to be represented in the respectively dimensioned model space. The arrangement of the individual parameters is then to be paid attention to. In principle, the mutual ordering of the individual parameters is arbitrary, it is true, but it should be ensured that a once selected arrangement is the same for all the speakers. More particularly, also the arrangement of the individual attribute parameters, which describe the individual probability densities of a certain state, is to be chosen for all the speakers, so that the parameters of all the speakers are correlated optimally. Only in the case of a similar arrangement of all the parameters in the supervectors of the individual speakers is it ensured that the determined basis vectors of the eigenspace correctly represent the desired information to differentiate various speakers.

Therefore, it is an object of the present invention to provide a method which ensures, when an eigenspace is determined, a similar arrangement of all the parameters in the supervectors of the individual speakers.

This object is achieved by a method as claimed in Claim 1.

The central idea of the invention is that in a first step a common speaker-independent set of models for the training speakers is developed, while the training speech data of all the training speakers involved are used. All the training speech data are then used for training respective speaker-independent models for the various acoustic units. Subsequently, in a second step, the training speech data of the individual training speakers are used for adapting the speaker-independent set of models found to the respective training speakers. This adaptation may be effected, for example, with the usual methods such as MAP or MLLR. When the models of the common speaker-independent set of models are adapted to the models of the speaker-dependent sets of models of the individual speakers, the respective semantic contents of the speech data are known. This is a so-called supervised adaptation. With this adaptation it can be detected without any problem which model parameters of the models of the speaker-independent set of models are assigned to the individual model parameters of the respective models of the speaker-dependent sets of models, so that also an unambiguous mutual assignment of the parameters can be determined. The concatenation of the individual model parameters of the sets of models to the

supervectors is then effected in such manner that the model parameters of the models of the speaker-dependent sets of models, which are assigned to the same model parameters of the same model of the speaker-independent common sets of models, are also arranged at the respective identical positions of the respective supervectors.

Thus, also with an extremely high number of different model parameters, an unambiguous arrangement of the individual parameters in the supervectors is guaranteed.

The method is particularly suitable for developing eigenspaces for speech systems that work on the basis of the Hidden Markow Models mentioned in the introduction. In principle, such a method, however, may also be used with other models where a multitude of parameters are to be concatenated to supervectors in a systematic manner, to thus represent, for example, the different speakers as dots in a high-dimensional model space.

In a particularly preferred embodiment of the method according to the invention, for determining the basis vectors of the eigenspace, the high-dimensional model space is first reduced to a speaker subspace via a simple change of basis, in which subspace the supervectors of all the training speakers are arranged. The actual transformation for determining the eigenspace basis vectors is then performed in this speaker subspace. Subsequently, the eigenspace basis vectors found are retransformed into the model space in a simple manner. Such a simple change of basis is possible, for example, by a Gram-Schmidt orthonormalization of the supervectors themselves or, preferably, by such orthonormalization of the difference vectors of the supervectors to a chosen original vector. A mean supervector is then preferably used as an original vector. This is the supervector whose parameters are the respective mean values of the parameters of the individual speaker-dependent supervectors. Such a simple change of basis can be performed in suitable manner also in high-dimensional spaces on current computers without any problem. To represent n different speakers in this speaker subspace, the speaker subspace is to have a maximum dimension of $n-1$ i.e. the dimensions of the space, in which then the actual calculation of the basis vectors of the eigenspace is performed, are strongly reduced compared with the dimensions of the original model space, so that considerable computing speed and memory capacity is saved. In the case of the model space mentioned in the introduction, of about two million dimensions to be used for recognizing continuous speech, the computers available at present need to have such a reduction of the necessary computing capacity and of the necessary main memory locations to be able at all to determine eigenspace basis vectors according to the transformation method mentioned in the introduction.

The various basis vectors of the eigenspace are preferably arranged according to their importance for distinguishing different speakers. This provides the possibility of reducing the eigenspace for the use in the speech recognition system even more in that the least important basis vectors of the eigenspace, which contain only little information by which the speakers can be distinguished, are rejected. The dimension of the eigenspace used last in a speaker recognition will then be much smaller than the number of training speakers. Thus, few coordinates will suffice to characterize the individual speaker- dependent models of the training speakers in the thus provided a priori optimized eigenspace within the high-dimensional model space and to perform an adaptation to a new speaker. The number of the necessary coordinates is then only a fraction of the number of degrees of freedom of other adaptation methods such as, for example, MLLR. When the PCA method is implemented for determining the eigenvectors of the covariance matrix of the supervectors as eigenspace basis vectors, the evaluation of the eigenvectors may be effected based on the associated eigenvalues. Eigenvectors having higher eigenvalues are more important than eigenvectors having lower eigenvalues.

According to the invention such an eigenspace can be used in a method of speech recognition, in which first a set of basic models is adapted to a current speaker while using the eigenspace, which adaptation is made on account of already observed speech data to be recognized of this speaker.

There are various possibilities for this. Several of them are discussed in EP 0 984 429 A2 mentioned above. The adaptation of the basic model to the respective speakers then takes place in the manner that the adapted model finally lies within the eigenspace, i.e. can be represented as a linear combination of the various basis vectors of the eigenspace.

The simplest method is the direct projection of the basic model in the eigenspace. Such a projection finds exactly the point within the eigenspace which lies closest to the new speaker's basic model lying outside the eigenspace. Unfortunately, this method is too coarse. In addition, such a projection operation comes up only when there is sufficient input speech material for the new speaker, so that all acoustic units are represented at least once in the data. In many applications these conditions cannot be satisfied.

As an alternative there is the possibility of the method also proposed in the above document, which is the method of Maximum Likelihood Eigenvoice Decomposition (MLEVD). With this method a point in the eigenspace is found, which exactly represents the supervector that belongs to a Hidden Markow model set, which has the greatest probability of

being generated by the new speaker's speech. This particular technique of this method will not be further discussed here. Reference is made in this respect to EP 0 984 429 A2.

With the two methods, also for such speakers who have characteristics that differ very much from the various characteristics of the training speaker, only one model set in the eigenspace is produced. Since such speakers, however, are not optimally represented by a model set in the eigenspace of the training speakers, it is practical first to develop a model set in the eigenspace with a method according to the invention and utilize this set as a new basic model set for a further optimal adaptation by means of a MAP or MLLR method. Obviously, however, any other method may be used too to adapt the basic model set to the new speakers while utilizing the eigenspace.

With the method according to the invention a relatively fast, single-speaker adaptation is possible with the recognition of a continuous speech signal. The predominant computing cost is to be made only once for processing the training speech material and for finding the eigenspace, while this computing cost itself can be controlled without any problem with continuous speech recognition with a large vocabulary, because a basis change is made for reducing the parameter space to a subspace. In this manner the eigenspace can be efficiently formed without directly using the millions of parameters.

The invention will be further explained with reference to the appended drawing Figures based on an example of embodiment. The characteristic attributes discussed hereinafter and the attributes already described above may be of essence to the invention not only in said combinations, but also individually or in other combinations.

In these drawings:

Fig. 1 gives a diagrammatic representation of the order of the various steps of the method according to the invention for generating an eigenspace,

Fig. 2 is an illustration of the adaptation of a speaker-independent model to two different speakers.

The method according to the invention is first started in that the whole speech data material of the training speakers, in the present example of embodiment of 300 different speakers, is used for training a common speaker-independent set of models with different speaker-independent models SI for the various acoustic units.

Such a speaker-independent model SI for an acoustic unit is shown in a solid-line elliptical distribution in the first step. In reality, this is a model consisting of three states, which are described by a plurality of probability densities. These densities are again described by 33 acoustic attribute parameters, which are each the mean value of the probability densities in the attribute space. In the following example of embodiment, 16 mel-cepstral coefficients and their 16 first time derivatives are used as attributes. The second derivative of the energy i.e. of the 0th mel-cepstral coefficient is added as the 33rd attribute. Obviously, such a model may also be described by fewer parameters, or by even more, additional parameters, for example, the variances of the density distributions.

In a second step, these speaker-independent models SI are adapted to the individual speakers while the respective training material of the training speakers is used, i.e. speaker-dependent models SD are generated. In the example of embodiment shown in Fig. 1 the speaker-independent model SI is adapted to four different speakers.

In Fig. 2 the method is clarified a little more with reference to an example having only two training speakers S_1 , S_2 . The dots correspond to respectively occurred observations of a specific acoustic unit, which was spoken by the two training speakers S_1 , S_2 . This is a representation in an attribute space which here has only two dimensions for clarity. Customarily, however, a single observation is not described by two, but by a multitude – in the present example of embodiment as stated, 33 – different attribute parameters. The attribute space is therefore in the present actual example of embodiment not two-dimensional as shown in Fig. 2, but 33-dimensional.

As may be noticed, the individual observations of the specific acoustic unit for the two speakers S_1 , S_2 is spatially divided over wide areas while, in the case shown, for the two speakers S_1 , S_2 two local maximums have clearly been formed. In one model this acoustic unit cannot therefore be sufficiently well formed with a single Gaussian probability density, but a superpositioning of at least two probability densities is to be used to represent the two maximums in the real spatial distribution of the observations. In reality the distribution of the observations is usually even more diffuse, so that for a good modeling about 30 probability densities are superimposed.

To determine what density of a model for a certain acoustic unit of a certain speaker corresponds to what density of the respective model for the same acoustic unit of another speaker, first a common speaker-independent model is trained from all the training speech data. In fig. 2 this model has exactly two speaker-independent Gaussian probability densities $SID^{(1)}$, $SID^{(2)}$. In a next step this speaker-independent model is then adapted to the

two individual speakers S_1, S_2 , while the known training speech material of the two individual speakers S_1, S_2 can be used. This leads to a speaker-dependent model having two probability densities $SDD^{(1)}, SDD^{(2)}_1$ for the first speaker S_1 and to another speaker-dependent model having two probability densities $SDD^{(1)}_2, SDD^{(2)}_2$ for the second speaker S_2 .

Since the models were developed from the same speaker-independent start model, the assignment of the probability densities is clear; the probability densities $SDD^{(1)}, SDD^{(2)}_1, SDD^{(1)}_2, SDD^{(2)}_2$ of the two speakers S_1, S_2 , which densities were developed from the same probability density $SID^{(1)}, SID^{(2)}$ of the speaker-independent model correspond. In the simplified case shown in Fig. 2 this correct assignment can also be seen with the naked eye from the position of the local maximums of the distributions of the individual observations of the speakers S_1, S_2 . The problem becomes evident, however, when it is considered that with a real evaluation of training speech data, not two distributions in a two-dimensional space, but approximately 30 distributions in a 33-dimensional space can be assigned to one another, while the individual distributions of a state of a model are slightly overlapping.

The assignment method according to the invention achieves that for each of the speakers there is a clear parameter assignment of each individual density, of each individual state and each individual model. Based on this known ordering, all the parameters for all the speakers can then be concatenated to one supervector per speaker, while it is ensured that in all the supervectors for all the speakers the same order of the parameters is present. With the aid of these supervectors, each individual speaker can be represented exactly as a dot in the high-dimensional model space, in the present example of embodiment an approximately 2×10^6 -dimensional space. This model space contains all the information of the speaker variations during the training.

For effectively using the information for a later speech recognition, a reduction of the data set, more particularly, a reduction of the dimensions of the model space is necessary, without essential information being lost then. For this purpose, the eigenvoice method is used, in which a transformation is performed of the supervectors of the individual speakers to find the basis vectors of an eigenspace. With this transformation, reduction criteria are used, which are based on the mutual variability, for example, on the variance, of the vectors to be transformed. A possibility of the transformation is – as discussed in the introductory part – the Principal Component Analysis (PCA). Also other suitable methods such as the Linear Discriminant Analysis (LDA), the Factor Analysis (FA), the Independent Component Analysis (ICA) or the Singular Value Decomposition (SVD) can obviously be used.

In the following example of embodiment it is assumed that for finding the eigenspace basis vectors, a PCA transformation is performed i.e. the eigenvectors of a covariance matrix determined by means of the supervectors of the individual speakers and the associated eigenvalues are searched for. These eigenvectors then form the eigenspace basis vectors.

In the following detailed mathematical description of this method, the following notations are used:

n_p is the number of the model parameters to be adapted; in the present example of embodiment the dimension of the attribute space (i.e. the number of the acoustic attribute parameters) multiplied by the total number of probability densities

n_s is the number of training speakers by which the training speech data were generated; $n_s \ll n_p$

\mathcal{P} is the model space i.e. the space of all the model parameters n_p , in the present example of embodiment they are all the mean values of all the probability densities in the attribute space. (Linear structures are used here.) This is to be taken into account when parameters to be adapted are represented. If, for example, variances σ are adapted, $\log(\sigma)$ is presented as a good coordinate to avoid transformations to very small or even negative values of σ . \mathcal{P} has the structure of an affine euclidian space, which means that vectors of an n_p -dimensional vector space can be used to define translations into \mathcal{P} in a natural way. The elements of \mathcal{P} are simply underlined in the following. Linear mappings in this space are underlined twice.

\underline{R}_i is an element of \mathcal{P} (for example, a set of model parameters of a possible speaker), i.e. a supervector of a speaker; $i = 1$ to n_s .

Instead of the covariance matrix of the supervectors themselves, the covariance matrix is determined of the difference vectors \underline{D}_i of the supervectors to a "mean supervector" \underline{R}_M of all the speakers.

Starting point for this is the determination of the mean value \underline{R}_M for the supervectors of all the \underline{R}_i of all the speakers:

$$\underline{R}_M = \frac{1}{n_s} \sum_i \underline{R}_i \quad (1)$$

This mean supervector \underline{R}_M is, as shown in formula (1), a supervector of all the speakers averaged component-by-component and thus represents an average set of models of the individual speaker-dependent sets of models of the training speakers.

Subsequently, the deviations i.e. the difference vectors \underline{D}_i of the individual supervectors \underline{R}_i from this mean supervector \underline{R}_M are determined:

$$\underline{D}_i = \underline{R}_i - \underline{R}_M \quad (2)$$

All further computations now take place with these difference vectors \underline{D}_i .

Fig. 2 shows in the method step shown bottommost for the two respective paired speaker-dependent densities $SDD^{(1)}_1, SDD^{(1)}_2$ and $SDD^{(2)}_1, SDD^{(2)}_2$ a mean density $MD^{(1)}, MD^{(2)}$. They are the densities $MD^{(1)}, MD^{(2)}$ that have the same variance as the speaker-dependent densities $SDD^{(1)}_1, SDD^{(1)}_2, SDD^{(2)}_1, SDD^{(2)}_2$. The mean value of these mean densities $MD^{(1)}, MD^{(2)}$ is the mean value of the mean values of the respective individual densities $SDD^{(1)}_1, SDD^{(1)}_2$ and $SDD^{(2)}_1, SDD^{(2)}_2$ of the two speakers S_1, S_2 .

The covariance matrix \underline{K} of the difference vectors \underline{D}_i is obtained from the multiplication of the vectors \underline{D}_i as rows in an $n_s \times n_p$ matrix \underline{D} with its transformed \underline{D}^T :

$$\underline{K} = \underline{D}^T \underline{D} \quad (3)$$

The in general n_s eigenvectors $\underline{E}_1, \dots, \underline{E}_{n_s}$ of this covariance matrix \underline{K} having eigenvalue > 0 are the eigenspace basis vectors searched for as such. These PCA eigenvectors n_s are the main axes of the covariance matrix \underline{K} or the "lethargy tensor". The eigenvectors each time correspond to the axes along which the individual speakers are distinguished from each other. Since the covariance matrix \underline{K} was built from the difference vectors \underline{D}_i of the supervectors \underline{R}_i formed to become the mean supervector \underline{R}_M , the eigenvectors $\underline{E}_1, \dots, \underline{E}_{n_s}$ run through the mean supervector \underline{R}_M , which forms the origin of the eigenspace.

For two dimensions, the directions of the eigenvectors in the lower part of fig. 2 are illustrated. They run through the center of the mean densities $MD^{(1)}, MD^{(2)}$ formed by the two speaker-dependent densities $SDD^{(1)}_1, SDD^{(1)}_2$ and $SDD^{(2)}_1, SDD^{(2)}_2$ in the direction of connection of the two speaker-dependent densities $SDD^{(1)}_1, SDD^{(1)}_2$ and $SDD^{(2)}_1, SDD^{(2)}_2$, each belonging to a respective mean density $MD^{(1)}, MD^{(2)}$.

Since, however, such transformations as the computation of the eigenvectors of a covariance matrix in a substantially 2×10^6 dimensional space require extremely much

computational circuitry and an enormous main memory capacity (to store the vectors and matrices for the necessary computation operations), these transformations can hardly be realized with the computers available at present. Therefore, a further step is necessary to reduce the space before the actual calculation of the eigenspace basis vectors.

For this purpose, first an orthonormal basis is searched for, which covers a subspace (in the following also called speaker subspace) in the model space, in which subspace all the difference vectors \underline{D}_i are represented which belong to the individual speakers or their models, respectively. To find this orthonormal basis, a simple basis transformation is performed which requires relatively little computational circuitry. In the following example of embodiment the Gram-Schmidt orthonormalization method is chosen. Obviously, also another simple transformation method can be used for a change of basis, for example, a Löwdin transformation.

When the Gram-Schmidt orthonormalization method is executed, first one of the difference vectors, for example \underline{D}_1 , is used as a first Schmidt basis vector \underline{S}_1 of the orthonormal basis searched for and only normalized. Subsequently, the second difference vector \underline{D}_2 is orthonormalized to this first Schmidt-basis vector \underline{S}_1 in that this second difference vector \underline{D}_2 is first projected on the first found Schmidt-basis vector \underline{S}_1 and the component parallel with the first Schmidt-basis vector \underline{S}_1 of the second difference vector \underline{D}_2 is subtracted from the second difference vector \underline{D}_2 . The remaining component of the second difference vector \underline{D}_2 , which component is perpendicular to the first Schmidt-basis vector \underline{S}_1 , is then normalized and thus forms the second Schmidt-basis vector \underline{S}_2 . Accordingly is done with these further difference vectors \underline{D}_3 to \underline{D}_S , while first all the components parallel with the already existing Schmidt-basis vectors \underline{S}_i are subtracted and the component perpendicular thereto is standardized as a new Schmidt-basis vector \underline{S}_3 to \underline{S}_S .

Since also such an orthonormalization of 300 vectors in a dimensional space of about 2×10^6 cannot be performed without further measures because of the limited storage capacity in normal computers, in the concrete example of embodiment this orthonormalization is performed block by block. It is then assumed that the main memory of the computer is capable of simultaneously storing $2n$ -supervectors. The procedure is then as follows:

First the $2n$ -vectors $\underline{D}_{1..2n}$ are orthonormalized and their representation is stored in the new found basis $\underline{S}_{1..2n}$.

For each further block of n -vectors \underline{D}_i , first for each block of n orthonormalized Schmidt-basis vectors \underline{S}_j , which were already found, the projection of the \underline{D}_i on this \underline{S}_j is subtracted. The projection coefficients \underline{D}_i in the orthonormal basis found are then stored for the representation of the \underline{D}_i in the orthonormal basis. Subsequently, the rest, i.e. the perpendicular components are mutually orthonormalized. The newly found Schmidt-basis vectors \underline{S}_j of the orthonormal basis and the representation coefficients of the individual difference vectors \underline{D}_i in this basis are then again stored.

Such a Gram-Schmidt orthonormalization needs

$$n_p \left(\frac{3}{2} n_s (n_s - 1) + n_s \right) \approx \frac{3}{2} n_s^2 n_p \quad (4)$$

floating point operations. They are, for example, with 300 speakers and 1 million dimensions, about 10^{11} individual operations, which can be carried out in about one to two seconds CPU time.

Since the difference vectors \underline{D}_i are correlated via the mean supervector \underline{R}_M according to the formulae (1) and (2), they are linearly dependent. Accordingly, an orthonormal basis vector is needed less than there are training speakers. This corresponds to the example to illustrate that three dots in a three-dimensional space can always be represented in a common plane i.e. also here a maximum of a two-dimensional subspace is necessary for representing the three dots of the three-dimensional space. The saving of one dimension (since the degree of freedom for the information – here unessential anyway – of the position of the speaker relative to the absolute zero of the model space is saved) is a reason why in the present example of embodiment the difference vectors \underline{D}_i of the speakers and not the supervectors \underline{R}_i themselves are used for forming the speaker subspace and for computing the covariance matrix. Furthermore, the coordinate jump in the eigenspace would otherwise be included, which, however, does not form a practical contribution to the speaker adaptation.

If desired, the supervector of the originally created common speaker-independent model can furthermore be represented for all the training speakers in this new Schmidt orthonormal basis. In this case the basis is naturally increased by one dimension. The dimension of the subspace then corresponds to the number of speakers, because the common speaker-independent model is represented by its own supervector irrespective of the

supervectors of the individual speakers, and thus represents an additionally mixed speaker which has an extremely large variance with regard to the individual acoustic units.

In lieu of the matrix $\underline{\underline{D}}$ of the distance vectors \underline{D}_i in the complete model space, now the representations of the difference vectors \underline{D}_i can be combined within the Schmidt orthonormal basis of the speech subspace in rows for a matrix $\underline{\underline{\theta}}$. This matrix $\underline{\underline{\theta}}$ is an $n_s \times n_s$ matrix i.e. it has only 300 times 300 elements. In contrast, the matrix $\underline{\underline{D}}$ of the difference vectors \underline{D}_i in the original model space has 300 times about 2 million elements.

For the sought covariance matrix $\underline{\underline{K}}$ of the difference vectors \underline{D}_i then holds

$$\begin{aligned}\underline{\underline{K}} &= \underline{\underline{D}}^T \underline{\underline{D}} \\ &= \underline{\underline{S}}^T \underline{\underline{\theta}}^T \underline{\underline{\theta}} \underline{\underline{S}}\end{aligned}\quad (5)$$

where $\underline{\underline{S}}$ is an $n_s \times n_p$ matrix of the basis vectors \underline{S}_i of the Schmidt orthonormal basis combined to columns. Since the basis vectors \underline{S}_i are orthonormal, a diagonalizing of $\underline{\underline{\theta}}^T \underline{\underline{\theta}}$ and a subsequent retransformation with the matrix $\underline{\underline{S}}$ is sufficient for finding the PCA eigenvectors $\underline{E}_1, \dots, \underline{E}_{n_s}$ in the model space. Since the vectors \underline{D}_i themselves have led to the orthonormalization, the matrix $\underline{\underline{\theta}}$ of the representations of the difference vectors \underline{D}_i in the Schmidt orthonormal basis is a triangular matrix, which renders the diagonalizing of $\underline{\underline{\theta}}^T \underline{\underline{\theta}}$ extremely simple.

The result is then an eigenspace whose dimension corresponds to the number of speakers -1, whose origin lies in the center of all original supervectors of the individual speakers and its basis vectors $\underline{E}_1, \dots, \underline{E}_{n_s}$ run along the variabilities of the individual speakers.

As an alternative, it is naturally also possible first to find an orthonormal basis of the supervectors themselves via a single change of basis, for example, a Gram-Schmidt orthonormalization. This basis found in this manner may then be shifted in the origin to the mean value of all the supervectors and, subsequently, the PCA method is executed first for determining the eigenvectors. This method of forming a Schmidt orthonormal basis from the supervectors themselves, a subsequent averaging in the new basis and a subsequent implementation of the PCA method as well as the subsequent retransformation, is shown in Fig. 1 in the last three method steps.

Naturally, the PCA method can also be executed with the orthonormal basis of the supervectors found by the simple change of basis and, subsequently, a transformation to a

desired origin. Furthermore, instead of the mean value of all the supervectors, also the supervector of the common speaker-independent model of all the training speakers may be used as the origin for the eigenspace.

The eigenspace found (and the representations of the speakers herein) is already considerably reduced compared to the original model space and still contains all information about the speaker variations in the training. However, it is still too complex to use during a rapid recognition. Therefore it is necessary for the dimension to be reduced more. This may be achieved in that simply several of the eigenvectors are rejected.

For this purpose, in the PCA method not only the eigenvectors, but also the associated eigenvalues of the covariance matrix \underline{K} may be determined. (eigenvalues are understood to mean in the sense of this document, unlike European patent application EP 0 984 429 A2 mentioned above, not the coefficients of a model when represented as a linear combination of the eigenvectors, but of the eigenvalue e belonging to the respective eigenvector \underline{E}_e of the matrix \underline{K} ; for which holds: $\underline{E}_e \underline{K} = e \underline{K}$). These eigenvalues may be used for determining an order of the eigenvectors \underline{E}_e . The higher the eigenvalue, the more important the associated eigenvector \underline{E}_e is for distinguishing between two different speakers. Therefore it is possible to select a certain number n_E of the most important eigenvectors, which are actually to be used for spreading out an eigenspace for a speech recognition system. In an example of embodiment of the method already implemented, they are only the eigenvectors having the ten largest eigenvalues, in another example the eigenvectors having the 50 most important eigenvalues.

It is self-evident that then only these eigenvectors actually used for spreading out the eigenspace, the so-called eigenvoices \underline{E}_e , are to be retransformed into the model space and not all the eigenvectors found of the covariance matrix \underline{K} . By selecting the basis for the eigenspace, it is ensured that, if a supervector \underline{R} , \underline{R}_i is projected on the reduced eigenspace, with a projection of a supervector \underline{R}_i on the reduced eigenspace with only n_E dimensions, including the original supervector \underline{R}_i , the resulting mean square error will certainly be minimized.

The eigenspace found in this manner may be used in several ways to adapt a basic model to a new speaker in a suitable way and in the fastest possible way. From this point of view this eigenspace can also be used as a complete data set in various speech recognition systems which utilize a data in a different way for adapting a basic model to a

new speaker, which data set already contains all the essential information of the training speech data in pre-evaluated manner.

11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60
61
62
63
64
65
66
67
68
69
70
71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
1001
1002
1003
1004
1005
1006
1007
1008
1009
1010
1011
1012
1013
1014
1015
1016
1017
1018
1019
1020
1021
1022
1023
1024
1025
1026
1027
1028
1029
1030
1031
1032
1033
1034
1035
1036
1037
1038
1039
1040
1041
1042
1043
1044
1045
1046
1047
1048
1049
1050
1051
1052
1053
1054
1055
1056
1057
1058
1059
1060
1061
1062
1063
1064
1065
1066
1067
1068
1069
1070
1071
1072
1073
1074
1075
1076
1077
1078
1079
1080
1081
1082
1083
1084
1085
1086
1087
1088
1089
1090
1091
1092
1093
1094
1095
1096
1097
1098
1099
1100
1101
1102
1103
1104
1105
1106
1107
1108
1109
1110
1111
1112
1113
1114
1115
1116
1117
1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150
1151
1152
1153
1154
1155
1156
1157
1158
1159
1160
1161
1162
1163
1164
1165
1166
1167
1168
1169
1170
1171
1172
1173
1174
1175
1176
1177
1178
1179
1180
1181
1182
1183
1184
1185
1186
1187
1188
1189
1190
1191
1192
1193
1194
1195
1196
1197
1198
1199
1200
1201
1202
1203
1204
1205
1206
1207
1208
1209
1210
1211
1212
1213
1214
1215
1216
1217
1218
1219
1220
1221
1222
1223
1224
1225
1226
1227
1228
1229
1230
1231
1232
1233
1234
1235
1236
1237
1238
1239
1240
1241
1242
1243
1244
1245
1246
1247
1248
1249
1250
1251
1252
1253
1254
1255
1256
1257
1258
1259
1260
1261
1262
1263
1264
1265
1266
1267
1268
1269
1270
1271
1272
1273
1274
1275
1276
1277
1278
1279
1280
1281
1282
1283
1284
1285
1286
1287
1288
1289
1290
1291
1292
1293
1294
1295
1296
1297
1298
1299
1300
1301
1302
1303
1304
1305
1306
1307
1308
1309
1310
1311
1312
1313
1314
1315
1316
1317
1318
1319
1320
1321
1322
1323
1324
1325
1326
1327
1328
1329
1330
1331
1332
1333
1334
1335
1336
1337
1338
1339
1340
1341
1342
1343
1344
1345
1346
1347
1348
1349
1350
1351
1352
1353
1354
1355
1356
1357
1358
1359
1360
1361
1362
1363
1364
1365
1366
1367
1368
1369
1370
1371
1372
1373
1374
1375
1376
1377
1378
1379
1380
1381
1382
1383
1384
1385
1386
1387
1388
1389
1390
1391
1392
1393
1394
1395
1396
1397
1398
1399
1400
1401
1402
1403
1404
1405
1406
1407
1408
1409
1410
1411
1412
1413
1414
1415
1416
1417
1418
1419
1420
1421
1422
1423
1424
1425
1426
1427
1428
1429
1430
1431
1432
1433
1434
1435
1436
1437
1438
1439
1440
1441
1442
1443
1444
1445
1446
1447
1448
1449
1450
1451
1452
1453
1454
1455
1456
1457
1458
1459
1460
1461
1462
1463
1464
1465
1466
1467
1468
1469
1470
1471
1472
1473
1474
1475
1476
1477
1478
1479
1480
1481
1482
1483
1484
1485
1486
1487
1488
1489
1490
1491
1492
1493
1494
1495
1496
1497
1498
1499
1500
1501
1502
1503
1504
1505
1506
1507
1508
1509
1510
1511
1512
1513
1514
1515
1516
1517
1518
1519
1520
1521
1522
1523
1524
1525
1526
1527
1528
1529
1530
1531
1532
1533
1534
1535
1536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565
1566
1567
1568
1569
1570
1571
1572
1573
1574
1575
1576
1577
1578
1579
1580
1581
1582
1583
1584
1585
1586
1587
1588
1589
1590
1591
1592
1593
1594
1595
1596
1597
1598
1599
1600
1601
1602
1603
1604
1605
1606
1607
1608
1609
1610
1611
1612
1613
1614
1615
1616
1617
1618
1619
1620
1621
1622
1623
1624
1625
1626
1627
1628
1629
1630
1631
1632
1633
1634
1635
1636
1637
1638
1639
1640
1641
1642
1643
1644
1645
1646
1647
1648
1649
1650
1651
1652
1653
1654
1655
1656
1657
1658
1659
1660
1661
1662
1663
1664
1665
1666
1667
1668
1669
1670
1671
1672
1673
1674
1675
1676
1677
1678
1679
1680
1681
1682
1683
1684
1685
1686
1687
1688
1689
1690
1691
1692
1693
1694
1695
1696
1697
1698
1699
1700
1701
1702
1703
1704
1705
1706
1707
1708
1709
1710
1711
1712
1713
1714
1715
1716
1717
1718
1719
1720
1721
1722
1723
1724
1725
1726
1727
1728
1729
1730
1731
1732
1733
1734
1735
1736
1737
1738
1739
1740
1741
1742
1743
1744
1745
1746
1747
1748
1749
1750
1751
1752
1753
1754
1755
1756
1757
1758
1759
1760
1761
1762
1763
1764
1765
1766
1767
1768
1769
1770
1771
1772
1773
1774
1775
1776
1777
1778
1779
1780
1781
1782
1783
1784
1785
1786
1787
1788
1789
1790
1791
1792
1793
1794
1795
1796
1797
1798
1799
1800
1801
1802
1803
1804
1805
1806
1807
1808
1809
1810
1811
1812
1813
1814
1815
1816
1817
1818
1819
1820
1821
1822
1823
1824
1825
1826
1827
1828
1829
1830
1831
1832
1833
1834
1835
1836
1837
1838
1839
1840
1841
1842
1843
1844
1845
1846
1847
1848
1849
1850
1851
1852
1853
1854
1855
1856
1857
1858
1859
1860
1861
1862
1863
1864
1865
1866
1867
1868
1869
1870
1871
1872
1873
1874
1875
1876
1877
1878
1879
1880
1881
1882
1883
1884
1885
1886
1887
1888
1889
1890
1891
1892
1893
1894
1895
1896
1897
1898
1899
1900
1901
1902
1903
1904
1905
1906
1907
1908
1909
1910
1911
1912
1913
1914
1915
1916
1917
1918
1919
1920
1921
1922
1923
1924
1925
1926
1927
1928
1929
1930
1931
1932
1933
1934
1935
1936
1937
1938
1939
1940
1941
1942
1943
1944
1945
1946
1947
1948
1949
1950
1951
1952
1953
1954
1955
1956
1957
1958
1959
1960
1961
1962
1963
1964
1965
1966
1967
1968
1969
1970
1971
1972
1973
1974
1975
1976
1977
1978
1979
1980
1981
1982
1983
1984
1985
1986
1987
1988
1989
1990
1991
1992
1993
1994
1995
1996
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
2016
2017
2018
2019
2020
2021
2022
2023
2024
2025
2026
2027
2028
2029
2030
2031
2032
2033
2034
2035
2036
2037
2038
2039
2040
2041
2042
2043
2044
2045
2046
2047
2048
2049
2050
2051
2052
2053
2054
2055
2056
2057
2058
2059
2060
2061
2062
2063
2064
2065
2066
2067
2068
2069
2070
2071
2072
2073
2074
2075
2076
2077
2078
2079
2080
2081
2082
2083
2084
2085
2086
2087
2088
2089
2090
2091
2092
2093
2094
2095
2096
2097
2098
2099
2100
2101
2102
2103
2104
2105
2106
2107
2108
2109
2110
2111
2112
2113
2114
2115
2116
2117
2118
2119
2120
2121
2122
2123
2124
2125
2126
2127
2128
2129
2130
2131
2132
2133
2134
2135
2136
2137
2138
2139
2140
2141
2142
2143
2144
2145
2146
2147
2148
2149
2150
2151
2152
2153
2154
2155
2156
2157
2158
2159
2160
2161
2162
2163
2164
2165
2166
2167
2168
2169
2170
2171
2172
2173
2174
2175
2176
2177
2178
2179
2180
2181
2182
2183
2184
2185
2186
2187
2188
2189
2190
2191
2192
2193
2194
2195
2196
2197
2198
2199
2200
2201
2202
2203
2204
2205
2206
2207
2208
2209
2210
2211
2212
2213
2214
2215
2216
2217
2218
2219
2220
22